DOCUMENT RESUME

ED 462 440 TM 033 704

AUTHOR McCamey, Randy
TITLE A Primer on the One-Parameter Rasch Model.
PUB DATE 2002-02-14
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (25th, Austin, TX, February 14-16, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Item Response Theory; Mathematical Models; Test Construction
IDENTIFIERS *One Parameter Model; *Rasch Model

ABSTRACT

The Rasch measurement model improves on traditional test construction by creating tests in which the person's ability is independent of the sample of items used and the norm group used to calibrate the test. This paper reviews the Rasch model by describing properties of the item characteristic curve (ICC) and discussing the utility of having person ability and item difficulty on a common scale. Calibrating item difficulty and person ability to a common scale allows test developers and instructors to create better instruments in terms of optimizing the number of items, eliminating items of the same difficulty, and more closely matching the level of difficulty of the items with the abilities of the examinees. (Contains 1 table, 2 figures, and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 440

Running Head: A primer on the Rasch Model

# A Primer on the One-Parameter Rasch Model

Randy McCamey

The University of North Texas

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

R. McCamey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033704

---

Paper presented at the 25th Annual Meeting of the Southwest Educational Research Association, February 14, 2002, Austin, TX.



BEST COPY AVAILABLE

Abstract

The Rasch measurement model improves on traditional test construction by creating tests in which the person's ability is independent of the sample of items used and the norm group used to calibrate the test. This article reviews the Rasch model by describing properties of the item characteristic curve (ICC) and discussing the utility of having person ability and item difficulty on a common scale.

## A Primer on the One-Parameter Rasch Model

Consider the following scenario: an employment test is created and subsequently standardized using the brightest, most intelligent workers in the company. In subsequent hiring situations, job applicants are tested and almost none pass the test, subsequently they are not hired. As a consequence, position's go unfilled, production falls behind, existing employees work overtime regularly, morale and productivity suffer and the company loses money. This scenario highlights one of the criticisms of traditional test construction – tests must be standardized or "norm'ed" correctly for the population being tested. If not, and the test is used outside the norm group parameters, the test results are essentially invalid and thus any decisions made using those test results become questionable. In specific hiring and selection processes, this scenario could also create, however unintended, adverse impact to one or more protected populations. Another criticism of traditional test construction dependent on norm groups is that, even if the test were standardized correctly, many populations change over time and thus old norms can become invalid for current applications of a test.

Beyond requirements of a norm group to standardize tests using traditional test construction methods, test often require a large number of items to measure a person's ability. If a method could be devised to provide a better assessment of test items such that items measuring the same ability level could be eliminated, then tests could be shortened. Test takers would be less likely to suffer from "test fatigue" as a result.

### Introduction to Rasch measurement

One modern model of measurement used in the social sciences is the 1-parameter item response theory (IRT) model. Georg Rasch, a Danish mathematician, had an interest

in teaching statistics and in measurement models, IRT models in particular. During the 1960's, Rasch developed his now-famous 1-parameter logistic model (the Rasch model) to estimate a person's trait level from their responses to test items (Embertson & Reise, 2000). Although the Rasch model and the 1-parameter IRT model use different algorithms for calculations, the results are virtually identical.

The Rasch model, as with IRT models in general, promised to overcome weaknesses in classical test theory. Specifically, IRT promises to overcome circular dependency of CTT which is the situation, as described by Fan (1998), where the person statistic is item dependent and the item statistic is examinee (person) dependent. The Rasch model improves on traditional test construction in the sense that Rasch creates item-free and person-free tests. That is, the Rasch model allows tests to be constructed where the measure of a person's ability is independent of the sample of items used and is independent of the norm group used to "calibrate" the test (Hashway, 1978). In the simplest form, a person's response to an item is the dependent variable in the Rasch model and the independent variables are the person's trait score (theta or $\theta$) and the item difficulty ($b$).

The Rasch model can be used for measurement (i.e., locating a person on the latent continuum) or exploratory data analysis (i.e., understanding the structure of items or selecting a useful subset of items). The Rasch model permits identification of items or behaviors that are ordered (e.g., what are the sequence of skills one needs to become a computer programmer) and thus the variable unit measure has the same meaning across the scale (Andrich, 1988). IRT modeling also allows statistical adjustments in scores and thus the development of more meaningful comparisons (Hambleton, 2000).

Using item response theory allows two distinct advantages over simple classical test theory. First, it allows researchers to more accurately rank respondents in terms of their patterns of responses (Crocker & Algina, 1986; Hambleton, 1983). Although some researchers have argued that IRT does not produce scores necessarily different from classical test theory, IRT is maximized at the tails of the distribution (Fan, 1998). This is an important consideration when working with individuals who tend to score at either extreme of a distribution. Second, using IRT estimates will allow for the generalization of scores to both the population of interest and to future users, whereas classical test theory results will not generalize to future users.

One of the practical applications of IRT modeling is to diagnose test instruments (i.e, item or test analysis). Table 1 lists the partial output of a Rasch analysis of a graduate level mid-term exam (n = 39) using the RASCAL for Windows (1995) software by Assessment Systems Corporation. It should be noted that IRT methods require much larger data sets, however this data set is introduced for heuristic purposes.

---

Insert Table 1 about here

---

Item difficulty (*b*) is the main parameter of interest in the Rasch model and is defined as the position on the latent trait variable where it is expected the person has a 50% probability of answering the item correctly. Note that item numbers 16, 7, and 13 all have the same item difficulty. Also note that there is a substantial difference between item difficulty for questions 16, 7, and 13 (-2.740) and items of the next higher item difficulty, items 15 and 19 (-2.028). Having this knowledge allows the researcher to modify one or more of items 16, 7, and 13 to fill in the item difficulty gap between

–2.740 and –2.028 if so desired. If the researcher is confident that the range of person abilities is being adequately measured by the test, there is evidence through this analysis to remove 2 of these 3 items (16, 7, and 3) with the same item difficulty. This allows the instructor or researcher to measure the same range of person abilities using a single item at the difficulty level of -2.740. Item analysis can be continued in this example as items 15 and 19 also have the same item difficulty (-2.028), as do items 2, 10, and 24 (-1.269). As this example shows, IRT modeling software can provide a convenient method for researchers to optimize both the number and difficulty of items on a test or assessment.

## Assumptions of the Rasch measurement

The first assumption of the Rasch model is that there is only one latent dimension underlying the items. This assumption is called *unidimensionality*; the item pool should be unidimensional and measure a single latent trait. This factor is not a severe limitation of the method since one can easily eliminate items that appear to violate the assumption. Harvey and Hammer (1999) also report that unidimensionality can be overcome by dividing the instruments into subscales or factors for those instruments with available subscales such as the Myers-Briggs Type Indicator.

A second assumption of the Rasch model is the local independence of items. That is, items should not give information that could be used to answer any subsequent item. Statistically, local independence means that the items do not correlate with each other (i.e., the items are uncorrelated or have a Pearson *r* at or near zero). Embretson and Reise (2000) describe this concept statistically as being the probability of solving any item *i* where the outcome of that item is independent of any other item.

## Item Characteristic Curve

The Rasch model can be used to measure magnitudes of variables using a single continuum (Andrich, 1988). IRT measurements are often used with, but not limited to, inventories or tests that utilize dichotomous responses (e.g., binary data, 0 and1). It is important to note that, in the Rasch model, items do not have to be dichotomous, merely the scoring of the items is required to be dichotomous in nature. Items may be any type that allows a yes/no or right/wrong scoring regardless of the number of possible choices (distracters) given. Several polytomous IRT models are available to handle multiple-ordered Likert-type responses such as the graded-response model, the partial credit model and the rating scale model (Harvey & Hammer, 1999; Embretson & Reise, 2000). To simplify the discussion, the remainder of this paper will assume dichotomous responses are used.

The Item Characteristic Curve (ICC) is a plot of the latent trait ($\theta$) on the x-axis by the probability of a correct response on the y-axis. The item characteristic function can be described as a mathematical representation of the relationship between a person's position on the latent trait dimension and the probability the person will correctly answer an item of a given difficulty (Hashway, 1978). The scale for the latent trait is typically described as a logarithmic measure (natural log or base e) thus forming a trait scale that is interval or near-interval in nature. According to Cella and Chang (2000), by creating an interval scale, parametric statistics are less subject to violation of assumptions and logit measures may temper bias at extremes in the scale. The Rasch model makes an assumption analogous to equal measurement error for each item and thus are said to be equally discriminating. The visual representation of this item characteristic function is the

item characteristic curve (ICC) as seen in Figure 1. This figure represents a plot of three separate ICC's on the same scale.

---

Insert Figure 1 about here

---

Again, item difficulty is defined as the point at which a person has a 50 percent probability of answering the item correctly. In Figure 1 for example, using three dichotomously scored items, the location where the item characteristic curves cross the 0.5 probability line is the item difficulty. Thus, item difficulty for the 3 items illustrated in Figure 1 is –1.0, 0.0, and 1.0 respectively. A function of the IRT is that it allows item difficulty and person ability to be plotted on the same scale. If a person's ability level exceeds an item's difficulty level, the person will generally pass the item (i.e., the probability of a correct answer increases) (Sarif, Cohen, & Costa, 1989).

## Item Response Theory Coefficients

The full IRT model produces 3 parameters for a given data set; parameters *a*, *b*, and *c*. Parameter *a* refers to the slope of the ICC. The slope tells the researcher something about the discriminating power of the item. As the slope becomes larger (i. e., more of a vertical orientation), the greater the ability of the item to discriminate between small changes in theta (person ability). As the slope of the ICC becomes less (i.e., more of a horizontal orientation), the less the ability of the item to discriminate between small changes in theta. In the 1-parameter model, parameter *a* is most often assumed to be 1.0 but may be fixed at some other predefined constant (Henson, 1999).

Parameter *c* is the "guessing parameter" and can help researchers take into account the respondent's ability to guess the answer to the item. For example, the probability of guessing a correct answer to a multiple-choice item with 4 options is 25

percent. In this case, at lower theta values, the ICC would become asymptotic to 0.25 rather than 0. In the Rasch model, parameter *c* is most often assumed to be 0 but may also be fixed at some other predefined constant (Henson, 1999).

## The Rasch Model

Common to the full IRT model and the Rasch model, parameter *b* is the item difficulty and is defined as the position on the latent trait variable where it is expected the person has a 50% probability of answering the item correctly. The further to the right on the plot the ICC stands, the greater the item difficulty as only those individuals with a higher theta would have a 0.5 or greater probability of having a correct answer. Figure 1 represents an ICC for 3 items of varying difficulty. An important concept to note is that theta ($\theta$), the latent trait or characteristic of the individual being measured, uses the same scale as the *b* parameter (item difficulty). According to Harvey and Hammer (1999), the location of the person and item parameters on a common scale represents an important, if not the critical, characteristic of IRT models. Hashway (1978) reinforces this concept as he discusses how the Rasch procedure assumes that both items and subjects occupy positions on the same latent trait dimension (i.e., the same scale).

Again, referring to Figure 1 for the Rasch model, the only characteristic distinguishing one item's difficulty from another is the location of the ICC on the horizontal axis (theta). The further left the ICC is on the graph, the lower the item difficulty and the further right the ICC is on the graph, the larger the item difficulty (i.e., the more difficult the item). The Rasch model also assumes that all Item Characteristic Curves are the same shape, which in the practical world is probably not completely true.

As noted previously, the Rasch model holds both the item discrimination parameter, *a*, and the guessing parameter, *c*, constant.

If the person's theta (latent trait) exceeds the item difficulty, the person is more likely to answer the item correctly. Conversely, if the person's theta is less than the item difficulty, the person will likely not answer the item correctly. This is an important point to the test developer. If the test item difficulty far exceeds the student's ability (theta), students will do poorly and the test will not yield significant information regarding the true ability of the students. Conversely, if the test item difficulty is significantly below that of the student's ability (theta), similar results occur: no significant information regarding student ability will be generated. IRT methods will often help discriminate between students with abilities at extremes of the distribution of scores by assisting the test developer in the development of items with many different item difficulties to assess different person (ability) levels.

For example, using three dichotomously scored items (see Figure 1), the location of subjects on the trait level continuum (x-axis) corresponds to their ability or trait level. The location of the items corresponds then to each item's difficulty levels. If a person's ability level exceeds an item's difficulty level, the person will generally pass the item (i.e., the probability of a correct answer increases) (Sarif et al. 1989).

Per the previous discussion, Hashway (1978) describes how the Rasch procedure places or calibrates both items and subjects (persons) to occupy positions on the same latent trait dimension (i.e., the same scale). Figure 2, an item by person distribution map generated from the same data set as that used for Table 1, is provided as a visual example to help relate the concept of trait level and item difficulty being on the same scale. Figure

2 shows the logit scale occupying the central potion of the map. Item difficulty (*b*) is displayed graphically to the left of the logit scale. Each marker (#) represents the percent of items at a particular item difficulty. Notice that several items, as a percentage, have an item difficulty of –2.8. From the previous discussion of item difficulty, using the values listed in Table 1, these markers correspond to items 16, 7, and 13. The map obviously rounds the item difficulty values. In this example, the item difficulty of –2.8 on the map corresponds to the calculated value of –2.740 for items 16, 7, and 13.

---

Insert Figure 2 about here

---

Person ability (theta) is displayed graphically on the right side of the logit scale on item by person distribution map (Figure 2). Again, each marker (#) represents the percent of examinees at a particular person ability or theta level. In this example, theta levels of the examinees reside at the upper level of the item difficulties. In some cases, theta levels exceed item difficulties.

In simple terms, Figure 2 shows the results of a Rasch assessment which creates a common scale for both item difficulty and person ability. In this example, it is easy to see that the item difficulties span a wide range (-2.8 to 2.0) and are generally below the person abilities. The person abilities are generally higher than the item difficulties and also in a narrower range (-0.2 to 3.4). What does this mean to the instructor or test developer? In general terms this assessment will allow the instructor to see the abilities of the students in relation to the difficulty of the items on a test. The logical outcome in this example is that the instructor or test developer could refine the test by removing items that have a low item difficulty, reducing the number of items that have the same item difficulty, and adding items at a higher difficulty level.

Summary

The Rasch model is gaining in use due to the widespread growth of computer applications and the increasing sophistication of computer programs to run demanding mathematical operations (Harvey & Hammer, 1999). The Rasch model assists test developers by providing a platform to calibrate instruments to be independent of the norm reference group. The Rasch model is also helpful in diagnosing instruments by calibrating item difficulty and person ability to a common scale. This function of the Rasch model allows test developers and instructors to create better instruments in terms of optimizing the number of items, eliminating items of the same difficulty, and more closely matching the level of difficulty of the items to the abilities of the examinees.

References

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications Inc.

Cella, D. and Chang, C. (2000). *A discussion of item response theory and its application in health status assessments.* Medical Care, 38(9), II-66-II-72.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart, and Winston.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists*. Mahway, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). *Item response theory and classical test theory: an empirical comparison of their item/person statistics*. Educational and Psychological Measurement, 58(3), 357-381.

Hambleton, R. K. (2000). *Emergence of item response modeling in instrument development and data analysis.* Medical Care, 38(9), II-60-II-65.

Hambleton, R. K. (1983). *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Harvey, R. J. and Hammer, A. L. (1999). *Item response theory*. The counseling psychologist, 27(3), 353-383.

Hashway, R. M. (1978). *Objective mental measurement: Individual and program evaluation using the Rasch model.* New York: Praeger Publishers.

Henson, R. K. (1999, January). *Understanding the one-parameter Rasch model of item response theory.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

RASCAL (Version 3.5) (1995). St. Paul, MN: Assessment Systems Corporation.

Safrit, M. J., Cohen A. S., and Costa M. G. (1989). *Item response theory and the measurement of motor behavior.* Research Quarterly for Exercise and Sport, 60, 325-335.

| Item Number | Item Difficulty |
|---|---|
| 16 | -2.740 |
| 7 | -2.740 |
| 13 | -2.740 |
| 15 | -2.028 |
| 19 | -2.028 |
| 2 | -1.269 |
| 10 | -1.269 |
| 24 | -1.269 |
| 14 | -1.006 |

Table 1.
*Item parameter estimates sorted by item difficulty*

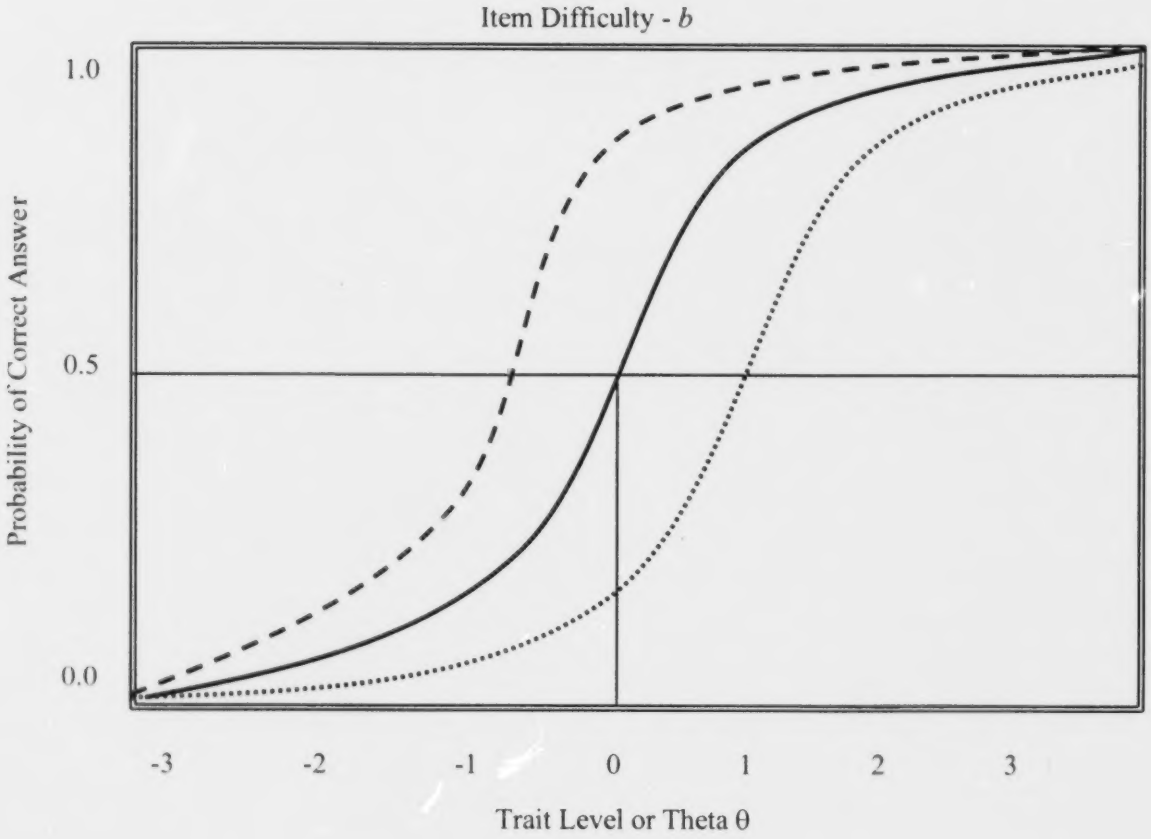


Figure 1.
*Item Characteristic Curves for 3 items of varying difficulty*

| Item Difficulty (*b*) | Logit Scale | Person Ability (Theta θ) |
|---|---|---|
| | -3.8 | |
| | -3.6 | |
| | -3.4 | |
| | -3.2 | |
| | -3.0 | |
| ##### | -2.8 | |
| | -2.6 | |
| | -2.4 | |
| | -2.2 | |
| ###+ | -2.0 | + |
| | -1.8 | |
| | -1.6 | |
| | -1.4 | |
| ##### | -1.2 | |
| #| | -1.0 | |
| | -0.8 | |
| ###| | -0.6 | |
| ###| | -0.4 | |
| ###| | -0.2 | |####
| + | 0.0 | +# |
| | 0.2 | |#######
| #| | 0.4 | |##
| #| | 0.6 | |####
| ###| | 0.8 | |######
| + | 1.0 | +## |
| | 1.2 | |
| | 1.4 | |######
| ########## | 1.6 | |
| ###| | 1.8 | |
| ###+ | 2.0 | +## |
| | 2.2 | |# |
| | 2.4 | |
| | 2.6 | |# |
| | 2.8 | |
| + | 3.0 | +####### |
| | 3.2 | |
| | 3.4 | |## |
| | 3.6 | |
| | 3.8 | |

Note. Data for this table was compiled from graduate level mid-term exams and is provided for illustrative purposes only.

Figure 2.
*Item by Person distribution map*